Lexicon Assistance Reduces Manual Verification of OCR Output

S. E. Hauser, A. C. Browne, G. R. Thoma, A.T. McCray
Lister Hill National Center for Biomedical Communications
National Library of Medicine, Bethesda, Maryland 20894, USA
Hauser@nlm.nih.gov, Browne@nlm.nih.gov

Abstract

An OCR system chosen for its high recognition rate and low percent of false positives also assigns low confidence values to many characters that are actually correct. Human operators must verify all words containing low confidence characters. We describe the creation of a lexicon optimized for automatically selectively resetting confidence values to high, thus reducing operator verification time. Two word lists, OCR Correct and OCR Incorrect, were extracted from files already processed and verified and became the standard for comparing candidate lexicons. A lexicon was selected from several candidate word lists maintained by the National Library of Medicine (NLM). In operation for about six months, lexicon assisted verification has been reducing the number of words requiring operator verification by over 50%.

Background

The Lister National Center for Biomedical Communications, a Research and Development Division of NLM, is developing a system [1] for semi-automated entry of journal article data into MEDLINE®, NLM's premier biomedical citation database used worldwide for clinical and research work. A first implementation of the data entry system has been in operation for about one year. In this system, the page or pages containing the text of each article's abstract are manually scanned, and the bitmapped image of the abstract is converted to a file of ASCII characters using Optical Character Recognition (OCR) technology. The OCR output is then manually verified and modified to conform to MEDLINE formatting standards. In addition to the ASCII characters obtained from the abstract image, the OCR output file also contains information about each character, including confidence level and character attributes. The confidence level, in the range from 0 to 9, indicates how certain the OCR software is that the given ASCII character is correct. During manual verification, the operator views the bitmapped image at the top of the computer screen and the ASCII equivalent at the bottom of the screen. Any letter that does not have the maximum confidence level of 9 is highlighted on the screen display, thus drawing operator attention to a word that may require correction.

The OCR server that is used by this system includes five separate commercial OCR engines, each of which processes the image. The server software employs a voting algorithm to select the correct ASCII character and assign its confidence value. The OCR server software was compared with five other commercial packages and was found to have a very high recognition rate and a very low number of incorrectly recognized characters with a high confidence value [2]. In tradeoff with the low percent of false positives, there are many low confidence characters that are correct. Consequently, verification operators frequently examine words that require no correction. If these words can be automatically verified by comparison with a lexicon, operator verification time should be reduced, and production level should increase.

Methods

A study was undertaken to create a lexicon of standard and medical words optimized for selectively un-highlighting words in OCR-generated abstracts for MEDLINE input. Words containing low confidence characters that are found in the lexicon would be un-highlighted, thus saving the verification operator the need to check those words. The ideal lexicon would contain all of the words that are correct irrespective of the confidence level assigned by the OCR system, while containing none of the words that are incorrect. Anticipating that a compromise would be necessary, the objective of the study was to select a lexicon that would maximize the number of correct words un-highlighted (benefit) while minimizing the number of incorrect words un-highlighted (cost).

The first step of the study was to extract a set of words containing low confidence characters from files already processed by the OCR system and divide them into lists of correct words and incorrect words. These two word lists, OCR Correct and OCR Incorrect, would become the standard for comparing candidate lexicons. The words for this study came from twenty-five journals randomly selected from those for which OCR conversion and verification had been completed. These contained a total of 565 abstracts, for a total of 139,958 words. A search of the initial output from OCR conversion for these 565 abstracts yielded 8085 words containing low confidence characters. For each of these words, the corresponding lines from the initial output files from OCR conversion and from the operator-verified files were combined into a third file for easy comparison. An in-house program and human assistants then examined each word to determine whether it was correct after OCR conversion, or whether the verification operator had changed it. The 8085 words were divided into three sets, shown in Table I.

TABLE I

	Number of words	Comments
OCR Correct	5188	Should be un-highlighted
OCR Incorrect	504	Should not be un-highlighted
Others	2393	Not relevant to this study

The 2393 “other” words include: a) words that were removed by the verification operators, such as keywords and publisher information at the end of an abstract; b) superscripts and subscripts, which are modified by the verification operators to conform to MEDLINE standards; c) Greek letters and other symbols that are not recognized by the OCR conversion software, and are also manually modified to conform to MEDLINE standards; d) words containing only digits. Of the 5693 candidate words for the study, over 90% were correctly converted by the OCR process. Un-highlighting these words could therefore represent significant savings in operator time.

The NLM maintains several word lists for its various services. These are good candidate lexicons as they are likely to contain much of the biomedical vocabulary present in MEDLINE abstracts. Four such lists are described briefly in Table II and in detail in the following paragraphs. These four lists and their combinations were used in the study.

The SPECIALIST Lexicon is a large syntactic lexicon of medical and common English vocabulary items that is released yearly with the Unified Medical Language System (UMLS) Knowledge Sources [3]. Each lexical record contains information on a variety of syntactic properties for each lexical item including inflectional patterns, the forms of plurals for nouns, of principal parts for verbs and of comparative and superlative for adjectives and adverbs [4]. Lexical items may be multi-words like "cardiac arrest". The UMLS Knowledge Source Server provides online access to the lexicon [5]. The lexicon is continuously updated and expanded. The 1997 release, used in this study, contains over 84,000 lexical records, accounting for 155,759 different strings, which were broken into a list of 132,598 unique words.

The Metathesaurus is the central vocabulary component of the UMLS. It is a database of concepts based on combining terms from more than 30 source vocabularies. We extracted 197,805 words from the 679,747 unique lower cased strings in the 1997 release of the Metathesaurus.

The Automated Indexing Management System (AIMS) word list is a frequency list of words that appear in MEDLINE abstracts. We combined AIMS lists from 1990-93 and 1994-1997. We also extracted a word list from the 1997 Medical Subject Headings (MeSH) Chemical Names.

TABLE II

Word List	Abbreviation (for this paper)	Number of Unique Words	Derived from
AIMS	A	417,277	MEDLINE abstracts
SPECIALIST Lexicon	S	132,598	Syntactic lexicon distributed with UMLS
MeSH Chemical Names	C	116,381	MeSH
Metathesaurus	M	197,805	UMLS Metathesaurus

The candidate lexicons were matched against the OCR Correct and OCR Incorrect lists. Words in the OCR Correct list that were in a lexicon were words that would be correctly un-highlighted. Words in the OCR Incorrect list that were in a lexicon were those that would be incorrectly un-highlighted. For each lexicon, the correctly un-highlighted words are the *benefit* available from using that lexicon and the incorrectly un-highlighted words are the *cost* of using that lexicon.

Results

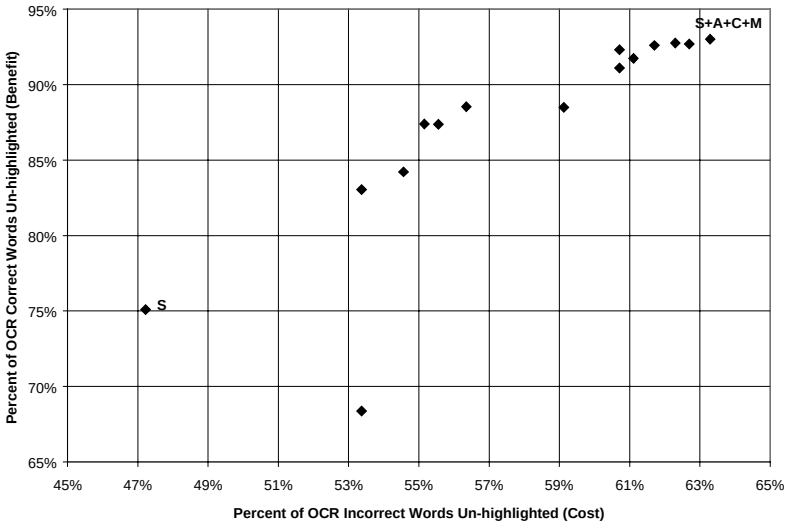

Figure 1. Benefit vs. cost for simple matching using individual and combined lexicons.

Simple matching of lists to lexicons yielded the poor results shown in Figure 1, where benefit is plotted against cost for individual and combined lexicons. In general, those lexicons that provided greater benefit also introduced greater cost. Even the lowest cost (47%) was too high to be considered. It was apparent that simple string matching would not be sufficient for OCR verification.

Inspection of the words containing low confidence characters revealed characteristics that could be used to eliminate certain words from lexicon checks to reduce cost more than benefit. Sixty-seven percent of the words in the OCR Incorrect list contain one or more characters with a confidence value less that 7, while only thirteen percent of the words in the OCR Correct list contain characters of confidence less than 7. Table

TABLE III

Words of Length ≤	% total words: OCR Incorrect list	% total words: OCR Correct list
1	23	7
2	41	23
3	59	41
4	74	60
5	85	68
6	90	74
7	94	79

III shows the cumulative distribution of word lengths in the two lists. There are more "short" words in the OCR Incorrect list than in the OCR Correct list. Based on these observations, rules using word length and confidence level were tested in conjunction with the lexicons in an effort to improve the cost-benefit ratio. Of the several rules tested, two resulted in reasonably low cost without proportionate reduction of benefit. The two rules are:

Rule 1: If the word
Is 6 or more characters long,
OR
Is 5 characters long and has no characters with confidence less than 7,
Then check for the word in the word list and, if found, un-highlight the word.

Rule 2: If the word
Is 6 or more characters long,
OR
Is 4 or 5 characters long and has no characters with confidence less than 7,
Then check for the word in the word list and, if found, un-highlight the word.

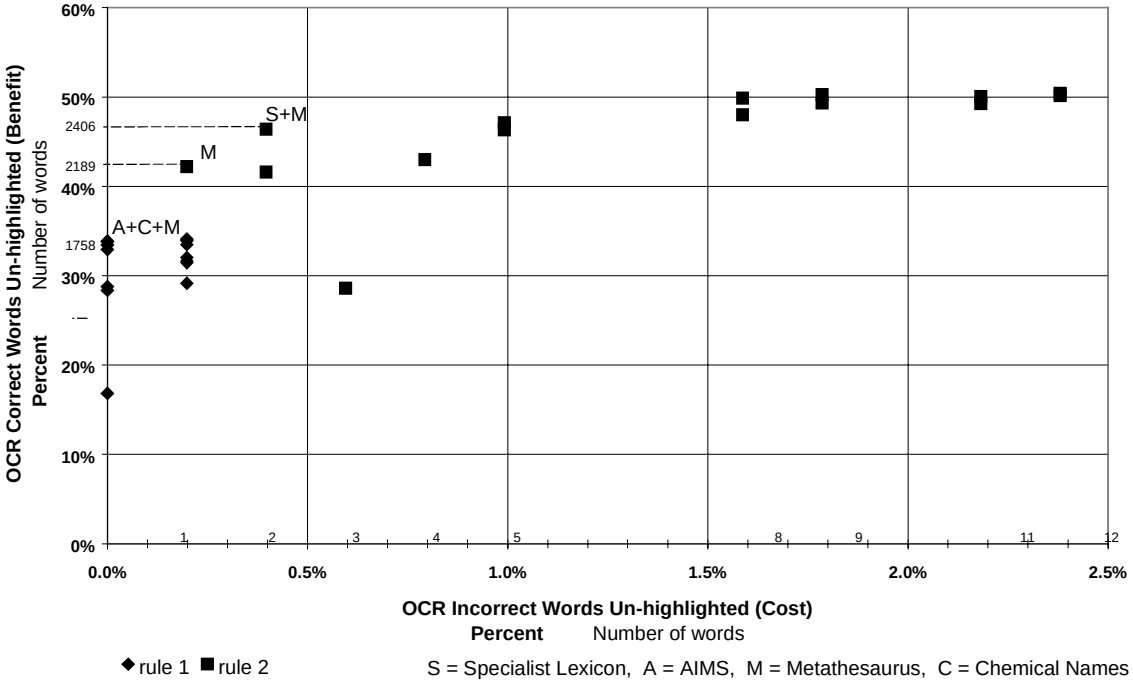

Figure 2. Benefit vs. cost using two word-length and confidence rules.

Figure 2 shows benefit vs. cost when these two rules are applied to all combinations of the four lexicons. The three best candidates are in the upper left portion of the graph. It is possible to have a system with no cost, and un-highlight about 33% of words that are correctly recognized by OCR. It is also possible to have a system with a cost less than 0.5% and un-highlight almost 50% of the correctly recognized words. Table IV summarizes the costs and benefits of the three best candidates for combination of rule and lexicon.

TABLE IV

Lexicon	Number of Unique Words	Rule	Percent OCR Correct Words Un-highlighted	Percent OCR Incorrect Words Un-highlighted
SPECIALIST Lexicon + Metathesaurus	262,798	2	46.4	0.4
Metathesaurus	197,805	2	42.2	0.2
AIMS + MeSH Chemical Names + Metathesaurus	296,632	1	33.9	0

The one word (0.2%) in the OCR Incorrect list that was in the Metathesaurus is "note", which was "mote" in the image. The two words (0.4%) in the OCR Incorrect list that were in the SPECIALIST Lexicon + Metathesaurus are "note" and "quipped", which was "equipped" in the image.

Implementation and production data

Based on the results of the study, we recommended using the SPECIALIST Lexicon together with the Metathesaurus lexicon in conjunction with rule 2 for assisting OCR verification. In addition to the favorable cost-to-benefit ratio found in the study, these lexicons are continuously revised and enhanced because they are UMLS knowledge sources. By employing the lexicons in this application, we enhance their value and extend their usefulness.

Our recommendation was adopted, and implemented in the production data entry system in early October of 1997. The selected word list was compressed and organized into a special dictionary format for fast searching using commercially available software [6]. An in-house C program, incorporating Dynamic Link Libraries from the same software product [6], parses words from the OCR output, and, depending on word length, confidence levels and attributes, checks for a match in the dictionary. If the word matches, confidence levels of all characters in the word are changed to 9; in other words, the word is un-highlighted. Our software implementation of lexicon assisted verification includes internal counts of the number of words highlighted before and after lexicon checking. A summary of the data thus produced since mid-October is shown in Figure 3. On average, lexicon checking has consistently reduced the highlighted words from approximately 13.5% to approximately 6.5%.

Conclusions and future directions

We have demonstrated that applying simple rules with an appropriate lexicon can significantly reduce the human labor required for OCR verification. Our success can be credited to the availability of biomedical-oriented lexicons, and to the meticulous creation of OCR Correct and OCR Incorrect lists for comparing the lexicons.

We will continue to explore methods for increasing the number of correct words that can be un-highlighted, as well as automatically correcting incorrect words. Results from a preliminary study suggest that we can un-highlight an additional 2.5% of the originally highlighted words by combining and checking words that are separated by an end-of-line hyphen. We intend to explore other researchers' use of transient dictionaries and heuristics [7], and two-character transformation [8]. We also plan to design a system to automatically recognize and compile candidate words to add to the lexicon.

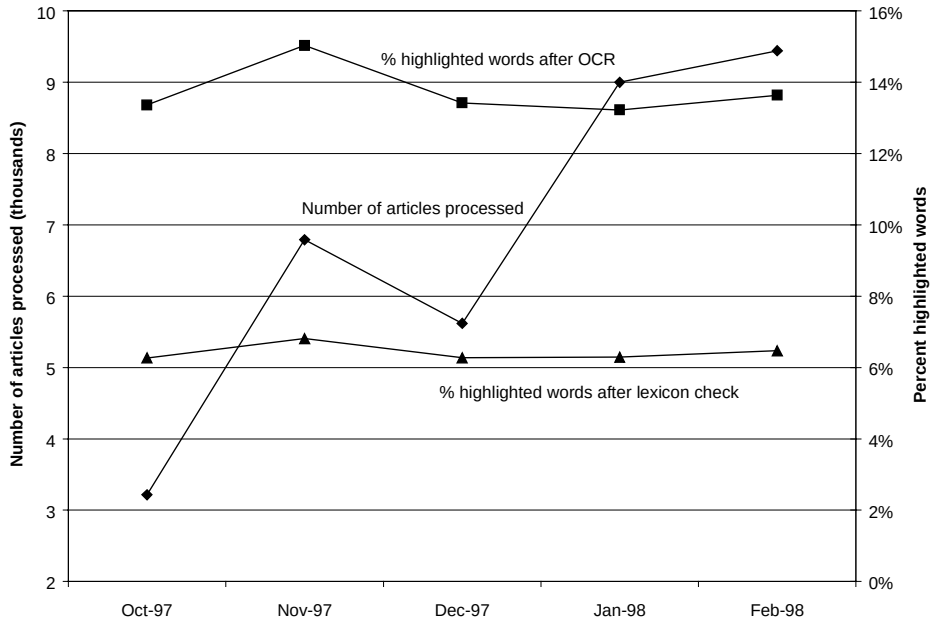

Figure 3. Results of lexicon checking on production data.

References

1. G.R. Thoma, D.X. Le. Medical database input using integrated OCR and document analysis and labeling technology. *Proceedings of the 1997 Symposium on Document Image Understanding Technology*. College Park, MD: University of Maryland, Institute for Advances Computer Studies; 1997; pp.180-1.
2. D.X. Le, V. Ngo, S. Wu. OCR test report. Internal document. National Library of Medicine, Bethesda, MD; 1996.
3. A.T. McCray, A.M. Razi, A.K. Bangalore, A.C. Browne, P.Z. Stavri. The UMLS knowledge source server: a versatile Internet-based research tool. J.J. Cimino (ed), *Proceedings of the 1996 AMIA Fall Symposium*; 1996; pp. 164-8.
4. A.T. McCray, S. Srinivasan, A.C. Browne. Lexical methods for managing variation in biomedical terminologies. J.G. Ozbolt (ed), *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*; 1994; pp. 235-9.
5. D.A.B. Lindberg, B.L. Humphreys, A.T. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, Vol. 32, No., 4; 1993; pp. 281-91.
6. Visual Speller version 1.01. From Visual Components, Inc.; 1995.
7. R.M.K. Sinha, B. Prasada, G.F. Houle, M Sabourin. Hybrid contextual text recognition with string matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 9; September 1993; pp. 915-25.
8. D.A. Dahl, L.M. Norton, S.L. Taylor. Improving OCR accuracy with linguistic knowledge. *Proceedings of the Second Annual Symposium on Document Analysis and information Retrieval*. University of Nevada, Las Vegas, NV; 1993; pp. 169-77.